New Results on DNA Codes

A. D'yachkov*, A. Macula†0, T. Renz†, P. Vilenkin* and I. Ismagilov*
*Department of Probability Theory
Faculty of Mechanics and Mathematics
Moscow State University, Moscow, 119992, Russia
Email: agd-msu@yandex.ru
†Air Force Res. Lab.
IFTC, Rome Research Site, Rome NY 13441, USA
Email: macula@geneseo.edu, thomas.renz@rl.af.mil

Abstract— **For q-ary n-sequences, we develop the concept of similarity functions that can be used (for $q = 4$) to model a thermodynamic similarity on DNA sequences. A similarity function is identified by the length of a longest common subsequence between two q-ary n-sequences. Codes based on similarity functions are called DNA codes [10]. DNA codes are important components in biomolecular computing and other biotechnical applications that employ DNA hybridization assays. We present our unpublished results [8] connected with the conventional deletion similarity function [1] used in the theory of error-correcting codes. The main aim of this paper – to obtain lower bounds on the rate of optimal DNA codes for a biologically motivated [11], [12], [13] similarity function called a similarity of blocks. We also present constructions of suboptimal DNA codes based on the parity-check code detecting one error in the Hamming metric [3].**

I. Introduction and Biological Motivation

Single strands of DNA are, abstractly, (A, C, G, T)-quaternary sequences, with the four letters denoting the respective nucleic acids. Strands of DNA sequence are oriented; for instance, $X = AACG$ is distinct from $Y = GCAA$. Furthermore, DNA is ordinarily double stranded: each sequence X, or strand, occurs with its *reverse complement* X', with reversal denoting that the sequences of the two strands are oppositely oriented, relative to one other, and with complementarity denoting that the allowed pairings of letters, opposing one another on the two strands, are (A, T) or (C, G)—the canonical Watson-Crick pairings. For instance, two sequences $X = AACG$ and $X' = CGTT$ are reverse complement of one another. Obviously, for any strand X, we have $(X')' = X$.

Whenever two, not necesseraly complementary, oppositely directed DNA strands "mirror" one another, they are capable of coalescing into a DNA duplex. The process of forming DNA duplexes from single strands is referred to as DNA *hybridization*. The greatest energy of DNA hybridization (the greatest stability of DNA duplex) is obtained when the two sequences are reverse complement of one another and the DNA duplex formed is a Watson-Crick (WC) duplex. However, there are many instances when the formation of non-WC duplexes are energetically favorable. The energy of DNA hybridization (the stability of DNA duplex) $\mathcal{E}(X, Y)$ of two single DNA strands X and Y is, to a first approximation, measured by the longest length of a common subsequence (not necessary contiguous) of either strand and the reverse complement of the other [10]. For two reverse complementary strands X and X' of length n, this measure plainly equals their length n, i.e., the maximum number of Watson-Crick bonds (complementary letter pairs) which may be formed between two oppositely oriented strands:

$$\mathcal{E}(X, X') = \max_{Y} \mathcal{E}(X, Y') =$$

$$= \max_{Y} \mathcal{E}(Y', X) = \mathcal{E}(X', X) = n. \qquad (1.1)$$

For instance, if $X = AACG$ and $X' = CGTT$, then $\mathcal{E}(X, X') = 4$.

A DNA code $\mathbf{X}$ is a collection of single stranded DNA sequences of fixed length n where each strand occurs with its reverse complement and no strand in the code equals its reverse complement [8], [10], i.e., if $X \in \mathbf{X}$, then $X' \in \mathbf{X}$ and $X' \neq X$. In DNA hybridization assays, the general rule is that formation of WC duplexes is good, but and the formation of non-WC duplexes is bad. A primary goal of DNA code design is to be assured that a fixed temperature can be found that is well above the melting point of all non-WC duplexes and well below the melting point of all WC duplexes that can form from strands in the code. Thus the formation of any WC duplex must be significantly more energetically favorable than all possible non-WC duplexes. DNA codes are important components for biomolecular computing [5] and other biotechnical applications that employ DNA hybridization assays. Note [10] that for these applications, the code length n, $10 \leq n \leq 40$, is experimentally accessible and that codes with more than 10^9 codewords could soon be called for.

The mathematical analysis of DNA hybridization is based on the concept of similarity functions that can be used to model a thermodynamic similarity on single stranded DNA sequences. For two quaternary n-sequences X and Y, the longest length of a sequence occurring as a (not necessary contiguous) subsequence of both is called a deletion similarity $S^{\lambda}(X, Y)$ between X and Y. We supposed [8], [10] that the deletion similarity $S^{\lambda}(X, Y)$ identifies the number of base pair bonds in a hybridization assay between X and the reverse complement of Y, i.e., the energy of DNA hybridization

[0] The work of A. Macula was supported by AFOSR – FA8750-05-C-0007

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **05 DEC 2005**	2. REPORT TYPE **N/A**	3. DATES COVERED -
4. TITLE AND SUBTITLE **New Results on DNA Codes**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Department of Probability Theory Faculty of Mechanics and Mathematics Moscow State University, Moscow, 119992, Russia**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR’S ACRONYM(S)
		11. SPONSOR/MONITOR’S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release, distribution unlimited**		
13. SUPPLEMENTARY NOTES **See also ADM001848, 2005 IEEE International Symposium on Information Theory Held in Adelaide, Australia on 4-9 September 2005.**		
14. ABSTRACT		
15. SUBJECT TERMS		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**5**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

$\mathcal{E}(X,Y')$ satisfying (1.1) is defined as follows

$$\mathcal{E}(X,Y') = \mathcal{E}(X',Y) = S^{\lambda}(X,Y) = S^{\lambda}(Y,X). \quad (1.2)$$

Let D, $1 \le D \le n-1$, be a fixed integer. A DNA code $\mathbf{X}$ is called a DNA code of distance D based on deletion similarity or, briefly, an (n,D)-code [8], [10] if the deletion similarity

$$S^{\lambda}(X,Y) \le n - D - 1, \qquad X,Y \in \mathbf{X}, \quad Y \ne X. \quad (1.3)$$

Definition (1.2) and condition (1.3) mean that the energy of DNA hybridization

$$\mathcal{E}(X,Y') \le n - D - 1, \qquad X,Y \in \mathbf{X}, \quad Y \ne X, \quad (1.4)$$

i.e., in (n,D)-code any strand X and the reverse complement of the other strand Y can never form $\ge n-D$ base pair bonds in a hybridization assay. In the theory of error-correcting codes, condition (1.3), by itself, specifies codes capable to correct any combination of D deletions [1], [4].

Example 1.1. DNA code $\mathbf{X} = \{X, X', Y, Y'\}$, where

$$X = ACAT, \quad X' = ATGT,$$
$$Y = ATAC, \quad Y' = GTAT, \quad (1.5)$$

is a (n,D)-code of length $n=4$ and distance $D=1$ because $n-D-1=2$ and sequence $Z = AT$ of length 2 is the longest common subsequence between any pair of strands in DNA code $\mathbf{X}$. Hence,

$$\mathcal{E}(X,X) = \mathcal{E}(X',X') = S^{\lambda}(X,X') = 2,$$
$$\mathcal{E}(Y,Y) = \mathcal{E}(Y',Y') = S^{\lambda}(Y,Y') = 2,$$
$$\mathcal{E}(X,Y) = \mathcal{E}(X',Y') = S^{\lambda}(X,Y') = 2,$$
$$\mathcal{E}(X,Y') = \mathcal{E}(X',Y) = S^{\lambda}(X,Y) = 2.$$

In papers [11], [12], [13], we introduced the concept of common block subsequence, namely: a common subsequence Z of sequences X and Y is called a common block subsequence if any two consecutive elements of Z which are consecutive in X are also consecutive in Y and vice versa. For two quaternary n-sequences X and Y, the longest length of a sequence occurring as a common block subsequence of both is called a block similarity between X and Y. For example, sequence $Z = AT$ of length 2 is the longest common block subsequence between any pair of strands in DNA code (1.5). Thus, DNA code (1.5) can be considered as DNA $(4,1)$-code based on block similarity.

The first conventional issue of coding theory [3] for DNA codes – to get a lower random coding bound on the rate of DNA codes and, hence, to identify values of the distance fraction D/n for which DNA code size grows exponentially when n increases. The given problem is more difficult than the corresponding problem for deletion-correcting codes. For instance, we cannot apply the best known random coding bounds [6] on the rate of deletion-correcting codes because these bounds were proved for codes which are not invariant under the reverse complement transformation. For the deletion similarity, the best known random coding bounds on the rate of DNA codes were established in our papers [8], [10]. The second conventional issue of coding theory for DNA codes – to present constructions of DNA codes. The aim of our paper is to obtain bounds and constructions for DNA codes based on the deletion and block similarities which have a good biological motivation to model a thermodynamic similarity on DNA sequences [11], [12], [13]. We will study q-ary DNA codes which can be considered as an evident generalization of quaternary DNA codes.

II. Notations, Definitions and Results

The symbol $\triangleq$ denotes definitional equalities and the symbol $[n] \triangleq \{1,2,\dots,n\}$ denotes the set of integers from 1 to n. Let $q = 2,4,\dots$ be a fixed even integer, $A \triangleq \{0,1,\dots,q-1\}$ be the standard alphabet of size $|A| = q$ and $\lfloor u \rfloor$ $(\lceil u \rceil)$ denote the largest (smallest) integer $\le u$ $(\ge u)$. Consider two arbitrary q-ary n-sequences

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \in A^n, \quad \mathbf{y} = (y_1, y_2, \dots, y_n) \in A^n.$$

In what follows, we will denote by symbol $S = S(\mathbf{x},\mathbf{y})$ an arbitrary symmetric function satisfying conditions

$$0 \le S(\mathbf{x},\mathbf{y}) = S(\mathbf{y},\mathbf{x}) \le S(\mathbf{x},\mathbf{x}) = n, \;\; \mathbf{x} \in A^n, \;\; \mathbf{y} \in A^n,$$

and called [10] a *similarity* function. Introduce the binary entropy function

$$h_q(u) \triangleq -u \log_q u - (1-u)\log_q(1-u), \quad 0 < u < 1.$$

Let $\ell \in [n]$ and $m = 1,2,\dots,\ell$. By symbol

$$\mathbf{z} = (z_1, z_2, \dots, z_\ell) \in A^\ell, \quad \text{where} \quad z_m = x_{i_m} = y_{j_m},$$
$$1 \le i_1 < i_2 < \dots < i_\ell \le n, \quad 1 \le j_1 < j_2 < \dots < j_\ell \le n,$$

we will denote a *common subsequence* of length $|\mathbf{z}| \triangleq \ell$ between $\mathbf{x}$ and $\mathbf{y}$.

Definition 1. [1]. Let $S^{\lambda}(\mathbf{x},\mathbf{y})$, $0 \le S^{\lambda}(\mathbf{x},\mathbf{y}) \le n$, denote the length $|\mathbf{z}|$ of *longest* common subsequence $\mathbf{z}$ between sequences $\mathbf{x}$ and $\mathbf{y}$. The number $S^{\lambda}(\mathbf{x},\mathbf{y})$ is called a *deletion similarity* between $\mathbf{x}$ and $\mathbf{y}$.

Definition 2. [11], [12], [13]. A common subsequence

$$\mathbf{z} = (z_1, z_2, \dots, z_\ell), \quad 2 \le \ell \le n,$$

is called a *common block subsequence* of length $|\mathbf{z}| \triangleq \ell$ between $\mathbf{x}$ and $\mathbf{y}$ if any two consecutive elements z_m, z_{m+1}, $m = 1,2,\dots,\ell-1$, which are consecutive (separated) in $\mathbf{x}$ are also consecutive (separated) in $\mathbf{y}$ and vice versa, i.e,

$$(z_m = x_{i_m}, z_{m+1} = x_{i_m+1}) \leftrightarrow (z_m = y_{j_m}, z_{m+1} = y_{j_m+1}).$$

Definition 3. [11], [12], [13]. Let $S^{\beta}(\mathbf{x},\mathbf{y})$ denote the length $|\mathbf{z}|$ of *longest* sequence occurring as a common block subsequence $\mathbf{z}$ between sequences $\mathbf{x}$ and $\mathbf{y}$. The number $S^{\beta}(\mathbf{x},\mathbf{y})$, $0 \le S^{\beta}(\mathbf{x},\mathbf{y}) \le n$, is called a *similarity of blocks* between $\mathbf{x}$ and $\mathbf{y}$. Obviously, $S^{\beta}(\mathbf{x},\mathbf{y}) \le S^{\lambda}(\mathbf{x},\mathbf{y})$.

Definition 4. [8], [10]. If $q = 2,4,\dots$, then

$$\bar{x} \triangleq (q-1) - x, \quad x \in A = \{0,1,\dots,q-1\},$$

is called a *complement* of a letter x. For an arbitrary q-ary n-sequence $\mathbf{x} = (x_1, x_2, \ldots, x_{n-1}, x_n) \in A^n$, we define its *reverse complement* $\tilde{\bar{\mathbf{x}}} \triangleq (\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1) \in A^n$.

Let $\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)$, where

$$\mathbf{x}(k) \triangleq (x_1(k), x_2(k), \ldots, x_n(k)),\ x_i(k) \in A,\ k \in [N],$$

be *codewords of a q-ary code* $\mathbf{X} = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)\}$ *of length n and even size N*. Let D, $1 \le D \le n-1$, be an arbitrary integer.

Definition 5. [8], [10]. A code $\mathbf{X}$ is called a DNA (n, D)*-code based on similarity function* $S = S(\mathbf{x}, \mathbf{y})$ (briefly, (n, D)-code) if the following two conditions are fulfilled. (i) For any number $k \in [N]$ there exists $k' \in [N]$, $k' \ne k$, such that $\mathbf{x}(k') = \tilde{\bar{\mathbf{x}}}(k)$. (ii) For any $k, k' \in [N]$, where $k \ne k'$, the similarity $S(\mathbf{x}(k), \mathbf{x}(k')) \le n - D - 1$. We will also say that code $\mathbf{X}$ is a DNA code of *length* n, *distance* D and *similarity* $n - D - 1$.

For $q = 4$, Definition 5 and a biological motivation of (n, D)-codes based on deletion similarity $S = S^{\lambda}(\mathbf{x}, \mathbf{y})$ were suggested in [10]. If only condition (ii) is retained, then an (n, D)-code based on deletion similarity is a code of length n capable to correct any combination of $\le D$ deletions [1]. A biological motivation of quaternary DNA codes based on similarity of blocks $S = S^{\beta}(\mathbf{x}, \mathbf{y})$ was suggested in [11].

For given n and D, we denote by $N_q(n, D)$ the *maximal size* of (n, D)-codes. If d, $0 < d < 1$, is a fixed number, then

$$R_q(d) \triangleq \overline{\lim_{n \to \infty}} \frac{\log_q N_q(n, \lfloor dn \rfloor)}{n}$$

is called a *rate* of $(n, \lfloor dn \rfloor)$-codes.

Let $d = d_q^{\lambda}$, $0 < d_q^{\lambda} < (q-1)/q$, be the unique root of equation $\frac{1+d}{2} = d \log_q(q-1) + h_q(d)$. A lower bound on the rate $R_q^{\lambda}(d)$ of DNA codes based on the deletion similarity is presented by

Theorem 1. [8]. *If* $0 < d < d_q^{\lambda}$, *then*

$$R_q^{\lambda}(d) \ge \underline{R}_q^{\lambda}(d) \triangleq 1 + d - 2[d \log_q(q-1) + h_q(d)].$$

Example 2.1. For the binary case, $d_2^{\lambda} = 0.13340$ and for the most important quaternary case, $d_4^{\lambda} = 0.27029$. In addition, $d_6^{\lambda} = 0.34902$ and $d_8^{\lambda} = 0.40324$.

Theorem 2. *For any distance fraction d,* $0 < d \le \frac{1}{2}$, *the rate* $R_q^{\beta}(d)$ *of DNA codes based on the similarity of blocks satisfies inequality*

$$R_q^{\beta}(d) \ge \underline{R}_q^{\beta}(d) \triangleq (1-d) - E_q(d), \tag{2.1}$$

$$E_q(d) \triangleq \max_{0 \le v \le d} F_q(v, d), \tag{2.2}$$

$$F_q(v, d) \triangleq (1-d)\, h_q\left(\frac{v}{1-d}\right) + 2d\, h_q\left(\frac{v}{d}\right). \tag{2.3}$$

Theorems 1 and 2 are established with the help of a *random coding bound* described in Sect. 3. The proof of Theorem 2 will be given in Sect. 4. The proof of Theorem 1 will be given in Sect. 5.

Let a number d_q^{β}, $0 < d_q^{\beta} \le 1/2$, be the unique root of equation $\underline{R}_q^{\beta}(d) = 0$ or $1 - d = E_q(d)$. Obviously, the lower bound $\underline{R}_q^{\beta}(d) > 0$ if $0 < d < d_q^{\beta}$ and we will say that d_q^{β} is a *critical point* of the lower bound $\underline{R}_q^{\beta}(d)$.

Example 2.2. We calculated $d_2^{\beta} = 0.17888$, $d_4^{\beta} = 0.35755$, $d_6^{\beta} = 0.44523$ and $d_8^{\beta} = 1/2$. It means that the critical points for block similarity exceed the corresponding critical points (see, Example 1) for deletion similarity.

One can easily understand that the conventional Hamming bound on the size of block codes with distance $D+1$ is a trivial upper bound on $N_q^{\beta}(n, D)$. For $D = 1$, an improvement of this trivial bound is given by

Theorem 3. *The maximal size* $N_q^{\beta}(n, 1) \le \left(q^{n-1} + q\right)/2$.

Proof. Consider an arbitrary code $\mathbf{X} = \{\mathbf{x}(k), k \in [N]\}$ of length n, distance $D = 1$ and block similarity $n - 2$. For each $\mathbf{x}(k)$, there exists one or two subsequences of length $n - 1$ obtained by *deletions of the first or the last element of* $\mathbf{x}(k)$. Let $\mathbf{X}$ contain N_1 (N_2) codewords which yield one (two) subsequences of length $n - 1$. Obviously, $N_1 \le q$. From item (ii) of Definition 5, it follows that there are $N_1 + 2N_2$ distinct $(n-1)$–subsequences, i.e., $N_1 + 2N_2 \le q^{n-1}$. Therefore, $N = N_1 + N_2 \le \left(q^{n-1} + q\right)/2$.

Theorem 3 is proved.

The following theorem is based on a construction of q-ary DNA codes obtained with the help of q-ary parity-check codes detecting one error in the Hamming metric [3].

Theorem 4. *There exists a q-ary DNA code of length n, distance* $D = 1$, *block similarity* $n - D - 1 = n - 2$ *and size*

- $N = \left(q^{n-1} + q\right)/2$ *if* $n = q = 2^m$;
- $N = q^{n-1}/2$ *if* $q = 2^m$, $n = 2^{m+k}$, $k \ge 1$;
- $N = \left(q^{n-1} - \frac{q^{n/2+1} - 1}{q-1}\right)/2$ *if n is a number divisible by q and* 4;
- $N = \left(q^{n-1} - q^{n/2} - \frac{q^{n/2+1} - 1}{q-1}\right)/2$ *if n is a number divisible by q.*

If $n = q = 2^m$, then Theorem 3 means that the construction of Theorem 4 is optimal. If q is fixed and $n \to \infty$, then Theorem 3 means that the construction of Theorem 4 is asymptotically optimal.

Example 2.3. If $q = 2$ and $n = 4$, then a DNA code of length $n = 4$, size $N = 4$, distance $D = 1$ and block (deletion) similarity $n - D - 1 = 2$ contains 2 pairs of codewords: **0000 1111** and **0110 1001**. Obviously, $N_2^{\beta}(4, 1) = N_2^{\lambda}(4, 1) = 4$.

Example 2.4. For $n = q = 4$, the construction of optimal DNA code from Theorem 4 is illustrated by the following table which contains $4^3 = 64$ codewords satisfying the parity-check condition: *for each codeword, the sum of its elements is a number divisible by* 4. These codewords are written as $\frac{1}{2} \cdot 4^3 = 32$ pairs of reverse complement codewords. Any row of the table consists of 1, 2, or 4 pairs. In any row, the first (second) codewords are obtained as consecutive cyclic shifts of the first (second) codeword of any fixed pair of the row. If we eliminate from the table all 15 pairs from the second and fourth columns of the table, then one can easily check that the rest 17 pairs will constitute a quaternary DNA code $\mathbf{X}$ of length $n = 4$, size $N = 2 \cdot 17 = 34$, block distance $D = 1$ and block similarity $n - D - 1 = 2$.

$\underline{\mathbf{0000}, \mathbf{3333}}$			
$\mathbf{0013}, \mathbf{0233}$	$\mathbf{3001}, \mathbf{2330}$	$\underline{\mathbf{1300}, \mathbf{3302}}$	$\mathbf{0130}, \mathbf{3023}$
$\underline{\mathbf{0022}, \mathbf{1133}}$	$\mathbf{2002}, \mathbf{1331}$	$\underline{\mathbf{2200}, \mathbf{3311}}$	$\mathbf{0220}, \mathbf{3113}$
$\mathbf{0031}, \mathbf{2033}$	$\mathbf{1003}, \mathbf{0332}$	$\mathbf{3100}, \mathbf{3320}$	$\mathbf{0310}, \mathbf{3203}$
$\underline{\mathbf{0103}, \mathbf{0323}}$	$\mathbf{3010}, \mathbf{3230}$	$\underline{\mathbf{0301}, \mathbf{2303}}$	$\mathbf{1030}, \mathbf{3032}$
$\mathbf{0112}, \mathbf{1223}$	$\mathbf{2011}, \mathbf{2231}$	$\underline{\mathbf{1201}, \mathbf{2312}}$	$\mathbf{1120}, \mathbf{3122}$
$\mathbf{0121}, \mathbf{2123}$	$\mathbf{1012}, \mathbf{1232}$	$\mathbf{2101}, \mathbf{2321}$	$\mathbf{1210}, \mathbf{3212}$
$\mathbf{0202}, \mathbf{1313}$	$\mathbf{2020}, \mathbf{3131}$		
$\underline{\mathbf{0211}, \mathbf{2213}}$	$\mathbf{1021}, \mathbf{2132}$	$\underline{\mathbf{1102}, \mathbf{1322}}$	$\mathbf{2110}, \mathbf{3221}$
$\underline{\mathbf{1111}, \mathbf{2222}}$			

We mark by the symbol *underline* pairs of codewords (there are 10 such pairs) from code X which have pairwise *deletion similarities* ≤ 2. They constitute a quaternary DNA code of length $n = 4$, size $N = 2 \cdot 10 = 20$, deletion distance $D = 1$ and deletion similarity $n - D - 1 = 2$. This means that the maximal size $N_4^{\lambda}(4, 1) \geq 20$. A general constructive lower bound on $N_4^{\lambda}(n, 1)$ is given by

Theorem 5. *If $n = qk$, where $k = 1, 3, \ldots$ is an odd number, then*

$$N_q^{\lambda}(n, 1) \geq \frac{q^{n-1}}{n}.$$

Theorem 5 is based on a construction [4] of codes correcting single deletions or insertions.

III. Random Coding Bound for DNA Codes

For an arbitrary integer s, $0 \leq s \leq n$, we define two sets

$$\mathcal{P}(n, s) \triangleq \{(\mathbf{x}, \mathbf{y}) \,:\, S(\mathbf{x}, \mathbf{y}) = s\}$$

and

$$\bar{\mathcal{P}}(n, s) \triangleq \{\mathbf{x} \,:\, S(\mathbf{x}, \tilde{\bar{\mathbf{x}}}) = s\},$$

i.e., the set of all pairs $(\mathbf{x}, \mathbf{y})$ for which the similarity $S(\mathbf{x}, \mathbf{y}) = s$ and the set of all sequences $\mathbf{x}$ for which the similarity between $\mathbf{x}$ and its reverse complement $\tilde{\bar{\mathbf{x}}}$ is $S(\mathbf{x}, \tilde{\bar{\mathbf{x}}}) = s$.

For fixed parameter u, $0 \leq u \leq 1$, define functions

$$\mathsf{p}(u) \triangleq \overline{\lim_{n \to \infty}} \frac{\log_q |\mathcal{P}(n, \lceil (1-u)n \rceil)|}{n}$$

and

$$\bar{\mathsf{p}}(u) \triangleq \overline{\lim_{n \to \infty}} \frac{\log_q |\bar{\mathcal{P}}(n, \lceil (1-u)n \rceil)|}{n}$$

satisfying inequalities $0 \leq \mathsf{p}(u) \leq 2$ and $0 \leq \bar{\mathsf{p}}(u) \leq 1$. One can obtain (the proof is omitted here) a random coding bound on the rate $R_q(d)$ which is given by

Lemma 3.1. *Let d, $0 < d < 1$, be fixed. If*

$$\min_{0 \leq u \leq d} \{1 - \bar{\mathsf{p}}(u)\} > 0,$$

then the rate

$$R_q(d) \geq \min_{0 \leq u \leq d} \{2 - \mathsf{p}(u)\}.$$

IV. Proof of Theorem 2

Let s, $1 \leq s \leq n$, be an arbitrary integer and $\mathcal{P}^{\beta}(n, s)$, $\bar{\mathcal{P}}^{\beta}(n, s)$ denote the sets from Lemma 3.1 for the similarity of blocks. For a fixed q-ary s-sequence $\mathbf{z} = (z_1, z_2, \ldots, z_s)$, and $j = 1, 2, \ldots, \min\{s, n - s + 1\}$, we introduce the concept of *j-block presentation* of $\mathbf{z}$, i.e., a *partition* of $\mathbf{z}$ into j *nonempty* blocks

$$\mathbf{z} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{j-1}, \mathbf{b}_j\}, \tag{4.1}$$

where each block contains *consecutive* elements of $\mathbf{z}$. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in A^n$, be a fixed q-ary n-sequence. We say that a block presentation $\mathbf{z}$ of the form (4.1) is a *block subsequence* (BSS) *of* $\mathbf{x}$ if $\mathbf{z}$ is a subsequence of $\mathbf{x}$, i.e.,

$$\mathbf{z} = \left(x_{i_1}, x_{i_2}, \ldots, x_{i_{s-1}}, x_{i_s}\right),$$

$$1 \leq i_1 < i_2 < \cdots < i_{s-1} < i_s \leq n,$$

and all blocks $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{j-1}, \mathbf{b}_j\}$ consisting of consecutive elements of the sequence $\mathbf{x}$ are *separated* in $\mathbf{x}$. Obviously, if a pair $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}^{\beta}(n, s)$ (a sequence $\mathbf{x} \in \bar{\mathcal{P}}^{\beta}(n, s)$), then there exists a block presentation $\mathbf{z}$ which is a *common BSS between* $\mathbf{x}$ *and* $\mathbf{y}$ ($\mathbf{x}$ and $\tilde{\bar{\mathbf{x}}}$), i.e., each of sequences $\mathbf{x}$ and $\mathbf{y}$ ($\mathbf{x}$ and $\tilde{\bar{\mathbf{x}}}$) contains separated blocks $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{j-1}, \mathbf{b}_j\}$ consisting of their consecutive elements. The following upper bound on the size $|\mathcal{P}^{\beta}(n, s)|$ is true.

Lemma 4.1. *For any s, $1 \leq s \leq n$, the size*

$$|\mathcal{P}^{\beta}(n, s)| \leq q^s \cdot$$

$$\cdot \sum_{j=1}^{\min\{s,\, n-s+1\}} \binom{s-1}{j-1} \cdot \left[q^{n-s} \cdot \binom{n-s+1}{j}\right]^2.$$

The proof of Lemma 4.1 is omitted here. For a fixed q-ary s-sequence $\mathbf{z} = (z_1, z_2, \ldots, z_s)$ and its j-block presentation (4.1), we introduce a *reverse complement* j-block presentation

$$\tilde{\bar{\mathbf{z}}} \triangleq \{\tilde{\bar{\mathbf{b}}}_j, \tilde{\bar{\mathbf{b}}}_{j-1}, \ldots, \tilde{\bar{\mathbf{b}}}_2, \tilde{\bar{\mathbf{b}}}_1\},\ j = 1, \ldots, \min\{s,\, n - s + 1\}.$$

Lemma 4.2. *The set $\bar{\mathcal{P}}^{\beta}(n, s)$ is empty if $s \geq 1$ is odd. If $s \geq 2$ is even and an n-sequence $\mathbf{x} \in \bar{\mathcal{P}}^{\beta}(n, s)$, then there exist an integer j, $j = 1, 2, \ldots, \min\{s,\, n - s + 1\}$ and a self-reverse complementary s-sequence $\mathbf{z} = \tilde{\bar{\mathbf{z}}}$, $|\mathbf{z}| = s$, of the form* (4.1) *which is a common block subsequence between* $\mathbf{x}$ *and* $\tilde{\bar{\mathbf{x}}}$ *and* $\mathbf{z}$ *has a self-reverse complementary block presentation*

$$\mathbf{z} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{j-1}, \mathbf{b}_j\} = \{\tilde{\bar{\mathbf{b}}}_j, \tilde{\bar{\mathbf{b}}}_{j-1}, \ldots, \tilde{\bar{\mathbf{b}}}_2, \tilde{\bar{\mathbf{b}}}_1\} = \tilde{\bar{\mathbf{z}}},$$

i.e., block $\mathbf{b}_1 = \tilde{\bar{\mathbf{b}}}_j$*, block* $\mathbf{b}_2 = \tilde{\bar{\mathbf{b}}}_{j-1}$*, ..., block* $\mathbf{b}_{j-1} = \tilde{\bar{\mathbf{b}}}_2$*, and block* $\mathbf{b}_j = \tilde{\bar{\mathbf{b}}}_1$.

The proof of Lemma 4.2 is omitted. Lemma 4.2 leads to

Lemma 4.3. *For any even s, $s \in [n]$, the size*

$$|\bar{\mathcal{P}}^{\beta}(n, s)| \leq q^{s/2} \cdot$$

$$\cdot \sum_{j=1}^{\min\{s,\, n-s+1\}} \binom{s/2 - 1}{\lceil j/2 \rceil - 1} \cdot \left[q^{n-s} \binom{n-s+1}{j}\right].$$

For $s \in [n]$, consider numbers

$$B(n,s) \triangleq \max_{1 \le j \le \min\{s,\, n-s+1\}} \left\{ \binom{s-1}{j-1} \cdot \binom{n-s+1}{j}^2 \right\}.$$

Let u, $0 < u < 1$, be fixed parameter. Introduce

$$E_q(u) \triangleq \lim_{n\to\infty} \frac{\log_q B\,(n,\,\lceil (1-u)n \rceil)}{n}, \qquad 0 < u < 1.$$

Lemmas 4.1 and 4.3 yield upper bounds on functions $\mathsf{p}^\beta(u)$ and $\bar{\mathsf{p}}^\beta(u)$ used in Lemma 3.1:

$$\mathsf{p}^\beta(u) \;\le\; (1+u) + E_q(u),$$

$$\bar{\mathsf{p}}^\beta(u) \;\le\; \frac{1}{2}\left[(1+u) + E_q(u)\right].$$

Therefore, Lemma 3.1 gives a random coding bound on the rate $R_q^\beta(d)$ of q-ary DNA $(n, \lfloor dn \rfloor)$-codes based on the similarity of blocks. One can easily check that the given lower bound $\underline{R}_q^\beta(d)$ can be written in the form (2.1)-(2.3).

Theorem 2 is proved.

V. Proof of Theorem 1

Let s, $\; 0 \le s \le n$, be an arbitrary integer and

$$\mathcal{P}^\lambda(n,s) \triangleq \{(\mathbf{x},\mathbf{y}) \,:\, S^\lambda(\mathbf{x},\mathbf{y}) = s\},$$

$$\bar{\mathcal{P}}^\lambda(n,s) \triangleq \{\mathbf{x} \,:\, S^\lambda(\mathbf{x},\tilde{\bar{\mathbf{x}}}) = s\},$$

denote the sets from Lemma 3.1 for the deletion similarity.

Lemma 5.1. [2], [7]. *Let n and s be integers, $0 \le s \le n$. For an arbitrary q-ary s-sequence $\mathbf{y}$ denote by $\mathbf{B}_q(\mathbf{y}, n)$ the set of all q-ary n-sequences $\mathbf{x}$ that include $\mathbf{y}$ as a subsequence, i.e., that can be obtained from $\mathbf{y}$ by $n-s$ insertions. Then for the fixed n and s, the size of $\mathbf{B}_q(\mathbf{y}, n)$ does not depend on $\mathbf{y}$ and has the form*

$$|\mathbf{B}_q(\mathbf{y},n)| = \sum_{k=0}^{n-s} \binom{n}{k} (q-1)^k \triangleq B_q(n,s). \tag{5.1}$$

Lemma 5.2 *The set $\bar{\mathcal{P}}^\lambda(n,s)$ is empty if s is odd. If s is even and an n-sequence $\mathbf{x} \in \bar{\mathcal{P}}^\lambda(n,s)$, then there exists a self-reverse complementary s-sequence $\mathbf{z} = \tilde{\bar{\mathbf{z}}}$, $|\mathbf{z}| = s$, which is a common subsequence between $\mathbf{x}$ and $\tilde{\bar{\mathbf{x}}}$.*

Lemma 5.2 is similar to Lemma 4.2. Lemmas 5.1 and 5.2 yield

$$|\mathcal{P}^\lambda(n,s)| \le\; q^s \,\cdot\, [B_q(n,s)]^2\,,$$

$$|\bar{\mathcal{P}}^\lambda(n,s)| \le\; q^{s/2} \,\cdot\, B_q(n,s), \quad 0 \le s \le n. \tag{5.2}$$

If u, $\; 0 \le u \le (q-1)/q$, is fixed, then from definition (5.1) it follows

$$\lim_{n\to\infty} \frac{\log_q B_q(n, \lceil (1-u)n \rceil)}{n} \;=\; u \log_q(q-1) + h_q(u).$$

Therefore, applying (5.2), we have

$$\mathsf{p}^\lambda(u) \triangleq \overline{\lim_{n\to\infty}} \; \frac{\log_q |\mathcal{P}^\lambda(n, \lceil (1-u)n \rceil)|}{n} \;\le$$

$$\le\; 1 - u + 2u \log_q(q-1) + 2h_q(u) \tag{5.3}$$

and

$$\bar{\mathsf{p}}^\lambda(u) \triangleq \overline{\lim_{n\to\infty}} \; \frac{\log_q |\bar{\mathcal{P}}^\lambda(n, \lceil (1-u)n \rceil)|}{n} \;\le$$

$$\le\; \frac{1}{2} \cdot [1 - u + 2u \log_q(q-1) + 2h_q(u)], \tag{5.4}$$

provided that $0 < u \le (q-1)/q$. Hence, if $0 < d < (q-1)/q$, then from (5.3)-(5.4) it follows

$$\min_{0 \le u \le d} \{2 - \mathsf{p}^\lambda(u)\} \ge$$

$$\ge 1 + d - 2d \log_q(q-1) - 2h_q(d) \triangleq \underline{R}_q^\lambda(d), \tag{5.5}$$

$$\min_{0 \le u \le d} \{1 - \bar{\mathsf{p}}^\lambda(u)\} \ge \frac{1}{2} \cdot \underline{R}_q^\lambda(d). \tag{5.6}$$

Inequalities (5.5)-(5.6) and Lemma 3.1 yield the statement of Theorem 1.

Theorem 1 is proved.

References

[1] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals", *Dokl. Akad. Nauk USSR*, vol. 163, pp. 845-848, 1965, (in Russian). English translation: *J. Soviet Phys.–Doklady*, **10**, pp. 707-710, (1966).

[2] V. I. Levenshtein, "Elements of Coding Theory", in the book: *Discrete Mathematics and Mathematical Problems of Cybernetics*, Moscow, "Nauka", 1974, pp. 207-305, (in Russian).

[3] F.J. MacWilliams, N.J.A. Sloane, *The Theory of Error - Correcting Codes*. Amsterdam, the Netherlands: North Holland, 1977.

[4] G.M. Tenengol'ts, "Nonbinary Codes, Correcting Single Deletions or Insertions", *IEEE Trans. Inform. Theory*, vol. IT-30, No 5, pp. 766-769, (1984).

[5] L. Adleman, "Molecular Computation of Solutions to Combinatorial Problems", *Science*, vol. 266, pp. 1021-1024, (1994).

[6] V. Dancik, "Expected Length of Longest Common Subsequence", Ph.D. thesis. On line: http://citeseer.nj.nec.com/

[7] V.I. Levenshtein, "Efficient Reconstruction of Sequences from Their Subsequences and Supersequences", *Journal of Combinatorial Theory*, Series A, vol.93, pp. 310-332 (2001).

[8] A.G. D'yachkov, D.C. Torney, P.A. Vilenkin, P.S. White, "On a Class of Codes for Insertion - Deletion Metric", *Proc. of ISIT-2002*, Lausanne, Switzerland, July 2002.

[9] V.I. Levenshtein, "Bounds for Deletion/Insertion Correcting Codes", *Proc. of ISIT-2002*, Lausanne, Switzerland, July 2002.

[10] D.C. Torney, A.G. D'yachkov, P.L. Erdos, V.V. Rykov, P.A. Vilenkin, P.S. White, "Exordium for DNA Codes", *Journal of Combinatorial Optimization*, 2003, v.7, pp. 369-379.

[11] A.J. Macula, A.G. D'yachkov, W.K. Pogozelski, T.E. Renz, V.V. Rykov, D.C. Torney, "An Insertion-Deletion Like Metric with Application to DNA Hybridization Thermodynamic Modeling", *IEEE Trans. Inform. Theory*, submitted.

[12] A.J. Macula, A.G. D'yachkov, W.K. Pogozelski, T.E. Renz, V.V. Rykov, D.C. Torney, "A Weighted Insertion-Deletion Stacked Pair Thermodynamic Metric for DNA Codes", *LNCS, Proc. of DNA 10*, Milan, Italy, 2004, to appear.

[13] A.J. Macula, A.G. D'yachkov, W.K. Pogozelski, T.E. Renz, V.V. Rykov, D.C. Torney, "New Insertion-Deletion Like Metrics for DNA Hybridization Thermodynamic Modeling", *Journal of Computational Biology*, submitted.